Curating a Pipeline for Analyzing and Visualizing Gene Expression Data in Psychiatric Disorders

Nicole Shedd (BCB/BBT)
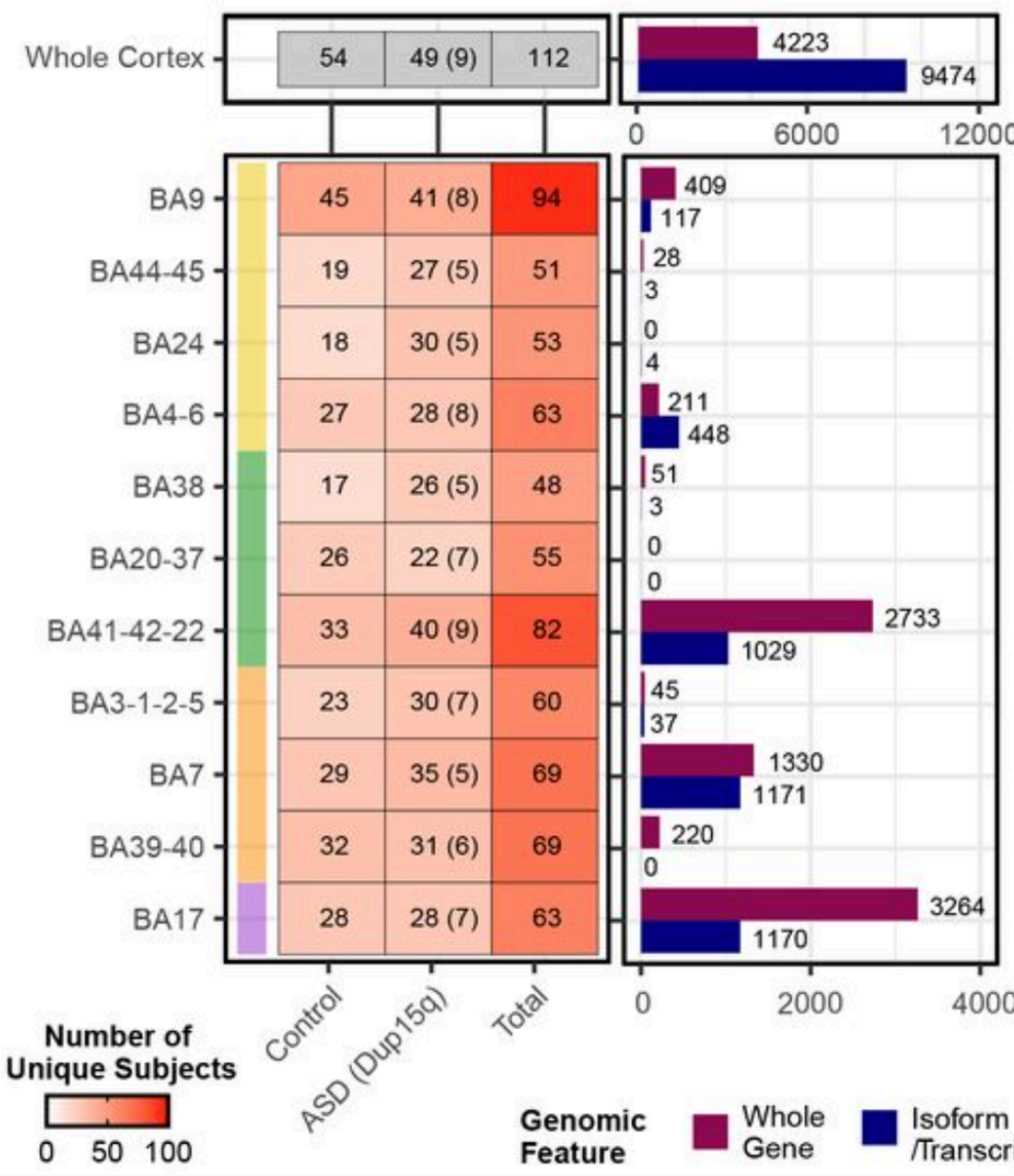Advisors: Elizabeth Ryder (BCB/BBT) and Inna Nechipurenko (BBT); Zhiping Weng and Henry Pratt (UMMS)

Genetics in Autism Spectrum Disorder (ASD)

Differentially expressed genes and isoforms from bulk RNA-seq data in ASD vary greatly by brain region:
- BA17 (primary visual cortex) has the most DEGs in this dataset
- BA 41-42-22 (primary auditory cortex) and BA7 (visuo-motor coordination cortex) followed
- BA4-6 (primary motor cortex) and BA9 (motor planning and organization) also had a significant number of DEGs

Differential gene expression aligns with clinical behavior associated with Autistic individuals:
- Visual and auditory perception differences are most frequent, followed by motor control

## Current Single-Cell Sequencing Methods

Single-cell sequencing allows us to understand cell-type specific changes to gene expression

Measuring gene expression:
RNA-seq: Uses next generation sequencing to find the quantity of RNA in a cell
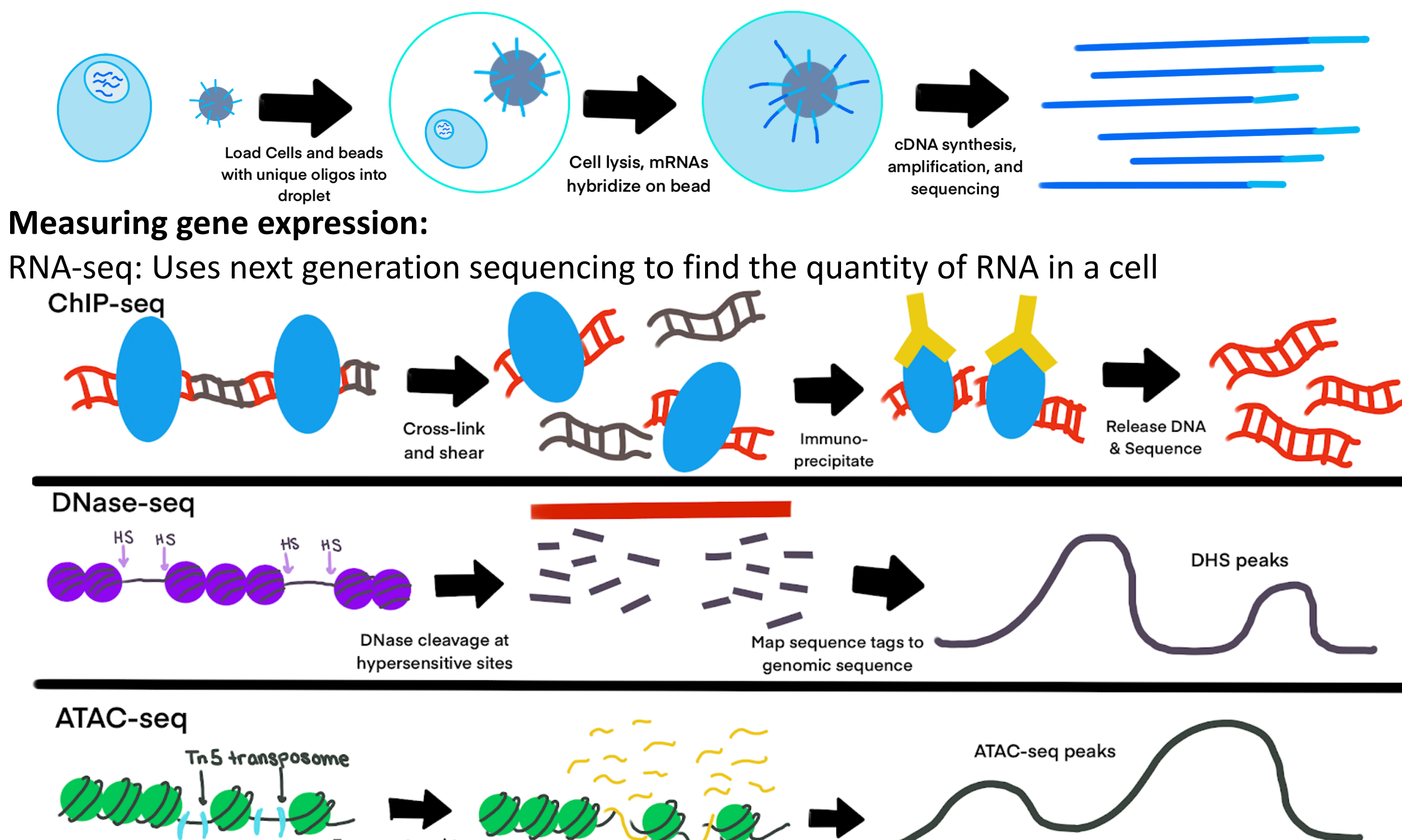
Measuring regulatory element enrichment:
ChIP-seq: locates DNA-binding sites for transcription factors and proteins using protein interactions
DNase-seq: maps DNase hypersensitive sites and identifies the cell's most active regulatory sites
ATAC-seq: probes DNA accessibility and maps transcription factor binding in accessible chromatin

Understanding Dimensionality Reduction
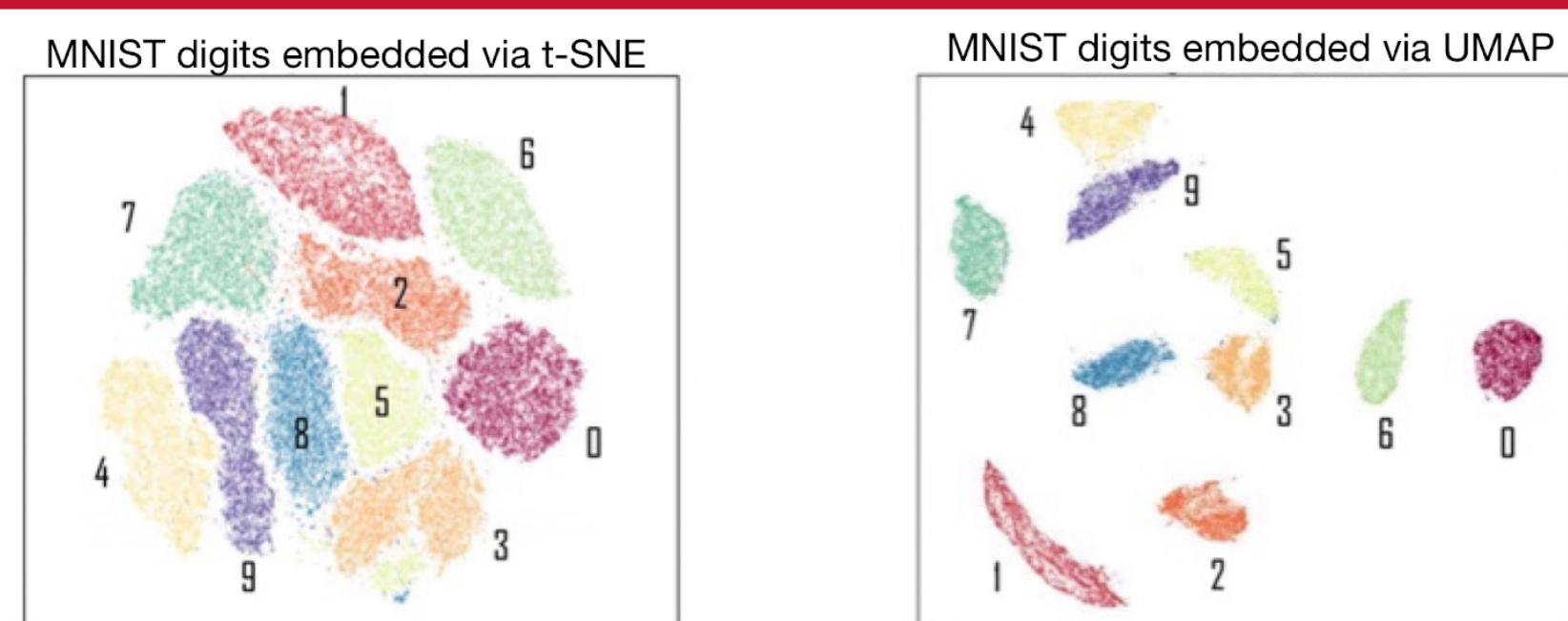
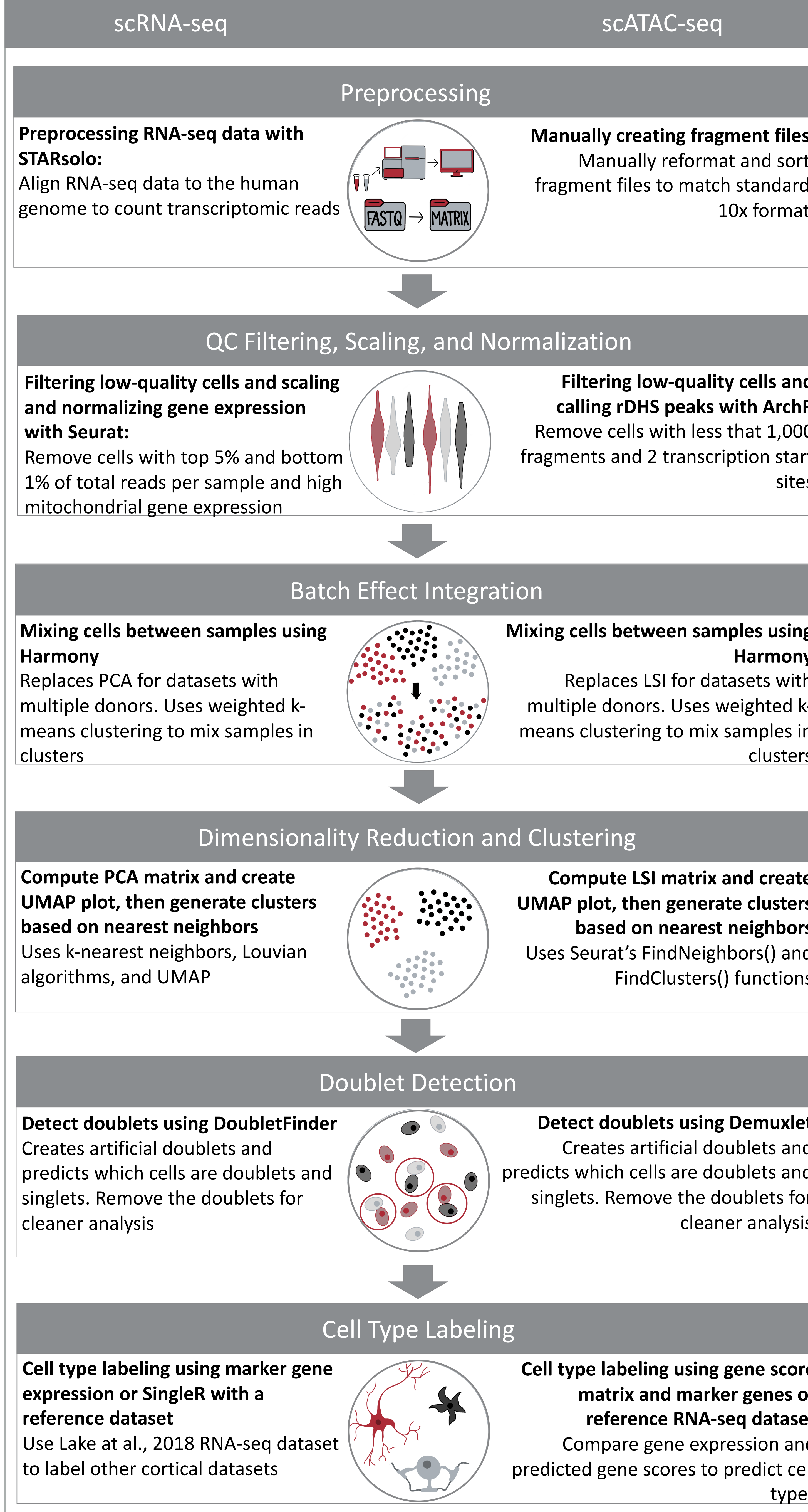**Plotting data on reduced dimensions: UMAP vs. t-SNE**
- t-SNE: Better for understanding similarity within clusters
- UMAP: better for understanding similarity between clusters

Visualization of 1500 handwritten digits using UMAP and t-SNE. In t-SNE, the digits 5 and 6 are located further apart than they are in UMAP. The digits look very similar, so when global structure is preserved, the clusters should appear closer.
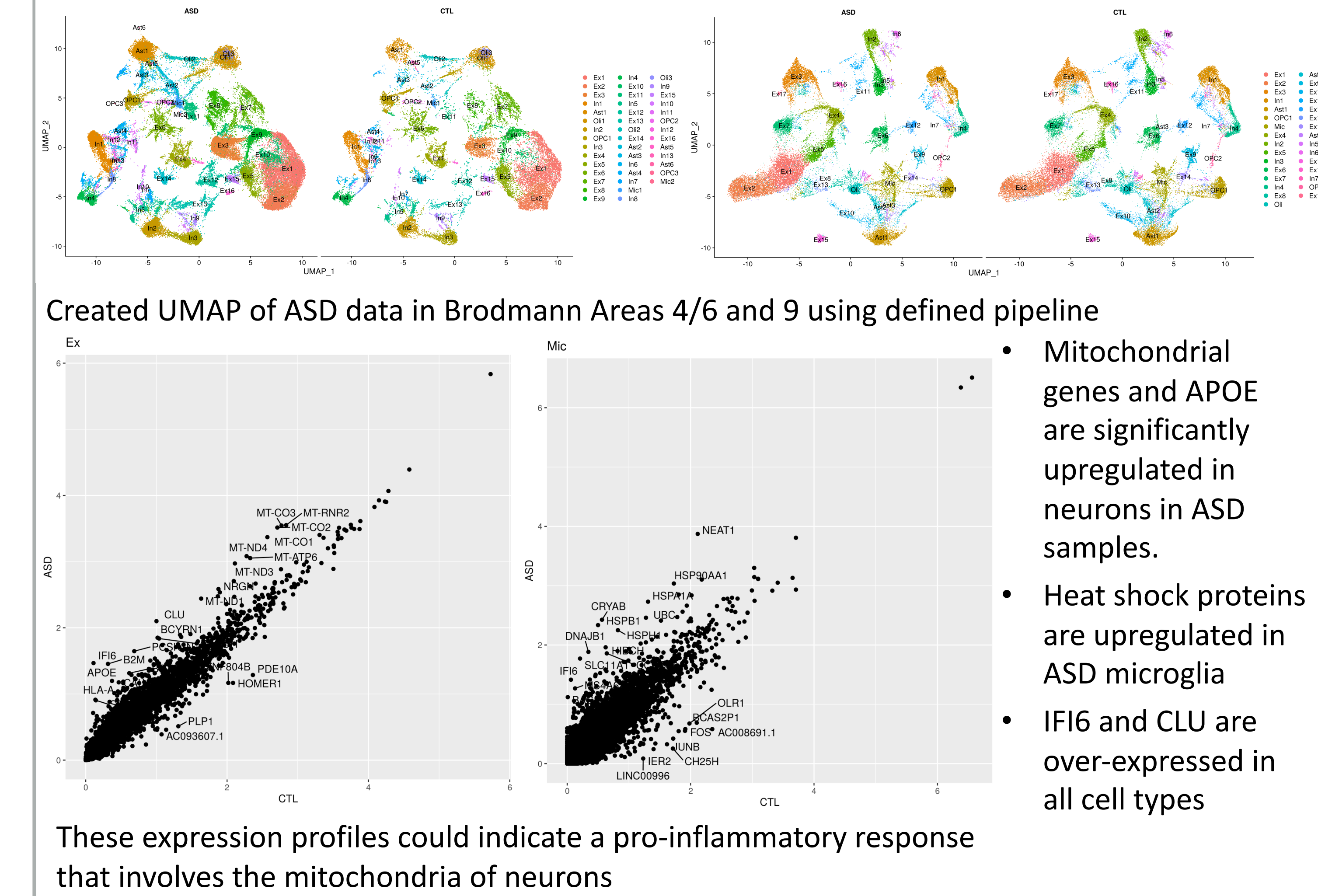
Objective

The goal of this project is to compile pipelines for analyzing single-cell RNA-seq and ATAC-seq data from the brain. I also applied this pipeline to understand cell-type specific changes to gene expression and cell type proportion in Autistic brains compared to non-Autistic controls.
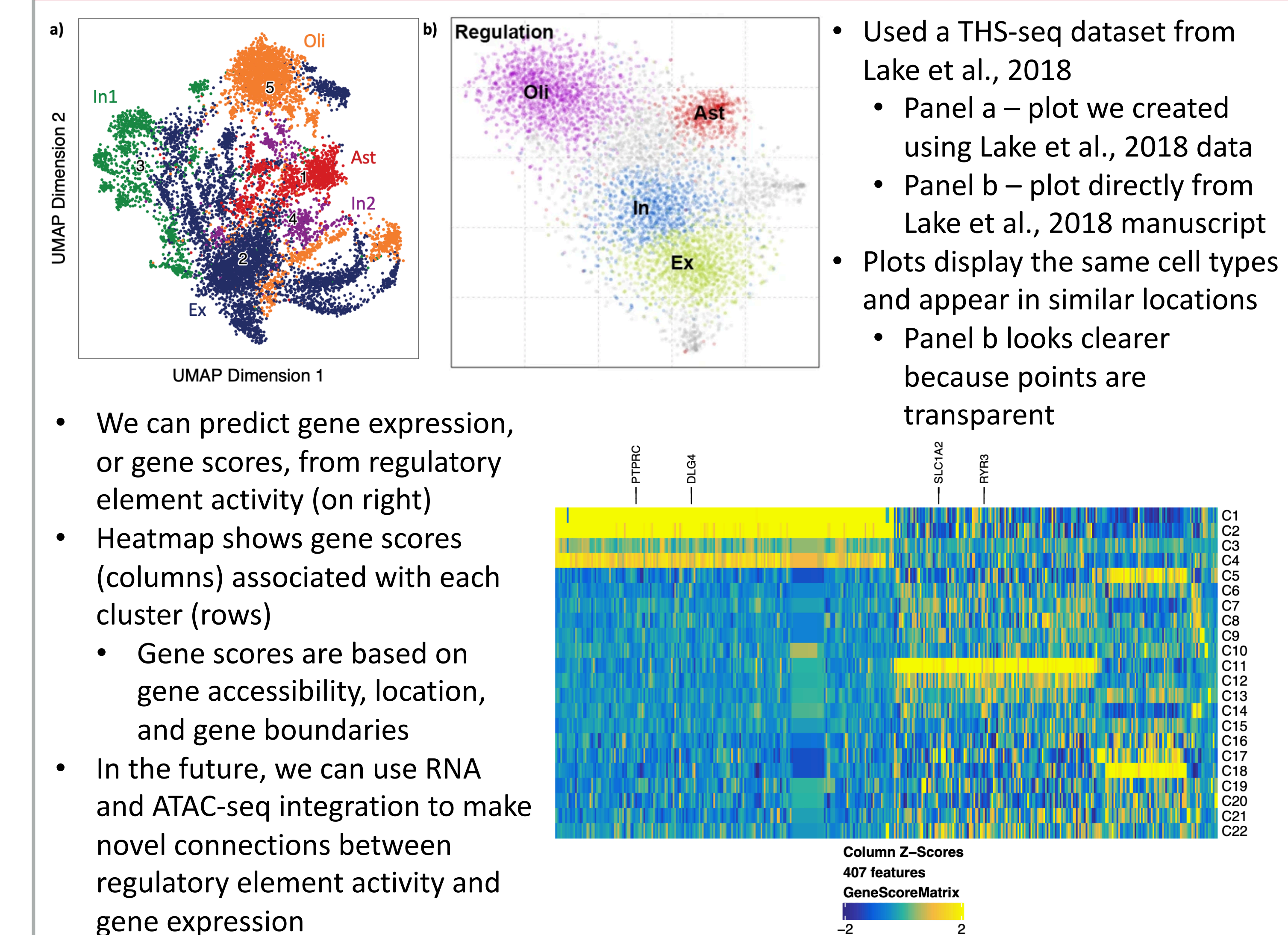
Proposed Analysis Pipeline

scRNA-seq	scATAC-seq

Preprocessing

Preprocessing RNA-seq data with STARsolo:
Align RNA-seq data to the human genome to count transcriptomic reads

Manually creating fragment files
Manually reformat and sort fragment files to match standard 10x format

QC Filtering, Scaling, and Normalization

Filtering low-quality cells and scaling and normalizing gene expression with Seurat:
Remove cells with top 5% and bottom 1% of total reads per sample and high mitochondrial gene expression

Filtering low-quality cells and calling rDHS peaks with ArchR
Remove cells with less that 1,000 fragments and 2 transcription start sites

Batch Effect Integration

Mixing cells between samples using Harmony
Replaces PCA for datasets with multiple donors. Uses weighted k-means clustering to mix samples in clusters

Mixing cells between samples using Harmony
Replaces LSI for datasets with multiple donors. Uses weighted k-means clustering to mix samples in clusters

Dimensionality Reduction and Clustering

Compute PCA matrix and create UMAP plot, then generate clusters based on nearest neighbors
Uses k-nearest neighbors, Louvian algorithms, and UMAP

Compute LSI matrix and create UMAP plot, then generate clusters based on nearest neighbors
Uses Seurat's FindNeighbors() and FindClusters() functions

Doublet Detection

Detect doublets using DoubletFinder
Creates artificial doublets and predicts which cells are doublets and singlets. Remove the doublets for cleaner analysis

Detect doublets using Demuxlet
Creates artificial doublets and predicts which cells are doublets and singlets. Remove the doublets for cleaner analysis

Cell Type Labeling

Cell type labeling using marker gene expression or SingleR with a reference dataset
Use Lake at al., 2018 RNA-seq dataset to label other cortical datasets

Cell type labeling using gene score matrix and marker genes or reference RNA-seq dataset
Compare gene expression and predicted gene scores to predict cell types

Application in Understanding ASD

Created UMAP of ASD data in Brodmann Areas 4/6 and 9 using defined pipeline

- Mitochondrial genes and APOE are significantly upregulated in neurons in ASD samples.
- Heat shock proteins are upregulated in ASD microglia
- IFI6 and CLU are over-expressed in all cell types

These expression profiles could indicate a pro-inflammatory response that involves the mitochondria of neurons

Validation of RNA-seq and ATAC-seq Integration

- Used a THS-seq dataset from Lake et al., 2018
  - Panel a – plot we created using Lake et al., 2018 data
  - Panel b – plot directly from Lake et al., 2018 manuscript
- Plots display the same cell types and appear in similar locations
- Panel b looks clearer because points are transparent

- We can predict gene expression, or gene scores, from regulatory element activity (on right)
- Heatmap shows gene scores (columns) associated with each cluster (rows)
  - Gene scores are based on gene accessibility, location, and gene boundaries
- In the future, we can use RNA and ATAC-seq integration to make novel connections between regulatory element activity and gene expression

Summary

- I compiled pipelines for analyzing single-cell sequencing data (both RNA-seq and ATAC-seq data)
- We validated both pipelines by using Lake et al., 2018 and comparing the final plots
- The plots looked very similar to the original publication, indicating that the pipeline will be successful

- Used the RNA-seq pipeline to understand gene expression in individuals with Autism Spectrum Disorder
- Large variations in cell type proportion by sample made differences insignificant
- Differential gene expression analysis supports neuroinflammation hypothesis involving mitochondrial dysfunction

Worcester Polytechnic Institute